Episode Book

EPISODE #6383

Choosing Class Breaks for Skewed Price Data

PUBLISHED	31 October 2026 at 15:45 UTC
DURATION	18:39
LISTEN	myweirdprompts.com/episode/choosing-class-breaks-pricing-tiers-32129/
HOSTS	Corn, Herman, Desmond, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	14 cited by the research (listed at the end)
BOOK GENERATED	1 November 2026 at 07:41 UTC

SYNOPSIS

Splitting continuous API prices into discrete tiers requires choosing between competing statistical philosophies. Cartography solved this problem decades ago through choropleth class interval selection, showing how variance-minimising methods like Jenks natural breaks contrast with quantiles, density cuts, and head-tail breaks. Preprocessing choices like logarithmic transforms, zero-price offsets, and goodness of variance fit often shape the resulting boundaries far more than the clustering algorithm itself.

TRANSCRIPT



Corn

Daniel has a column of batch API prices from OpenRouter, per million output tokens, and he can see the shape by eye. A very cheap cluster, a middle, a few expensive ones. He wants four tiers for a quadrant chart, and he wants the cut points to follow the real groupings instead of getting yanked around by a handful of extreme models.



Corn

He named Jenks, quantiles, one-dimensional k-means, Gaussian mixtures, density valleys, and log scales. What he wants to know is how you choose, and what's the established way.



Herman

Herman Poppleberry, and this one is a pleasure, because it's a solved problem in cartography. It's called class interval selection. People have been arguing about where to cut choropleth maps for decades.



Corn

Choropleth being the maps where regions are shaded by a value.



Herman

Right. Unemployment by county, rainfall by district. Every one of those maps forces the same decision Daniel has. You have a continuous number and you have to chop it into a handful of shades. Change where you chop, and the map tells a different story.



Corn

Give me a concrete case. How different can the story get?



Herman

Take a map of household income by county. Cut it into equal-width bands and one wealthy suburb county can turn nearly the whole country into the palest shade, with a single dark blob on the coast. Cut the same numbers into equal-count bands and the country looks evenly mixed, with dark patches everywhere. Same data, and two opposite impressions of how unequal the place is.



Herman

George Jenks was a cartographer at the University of Kansas. His worry was making maps visually honest. He talked about a blanket of error that should be spread evenly across the map instead of piling up in one class.



Corn

So a man worried about map shading wrote the algorithm Daniel's about to run on language model prices.



Herman

Yes. And here's the reframe. Daniel's list isn't five algorithms. Jenks and one-dimensional k-means are the same optimisation. Jenks is the cartographic name, k-means is the machine learning name. Wikipedia says it plainly, that Jenks "seems to be one dimensional k-means."



Corn

Well, that's one item off the menu for free.



Herman

What's left is four philosophies. Variance-minimising, which is Jenks and k-means. Count-equalising, which is quantiles. Density-valley, which is kernel density and Gaussian mixtures. And hierarchy, which is head and tail breaks. You're choosing a philosophy, not a library.



Corn

Then start with the variance one, since it's two for the price of one.



Herman

The objective is to make each class tight around its own mean while pushing the class means far apart. Minimise within-class variance, maximise between. You get a fit number out of it called goodness of variance fit. It's the total squared deviation around the overall mean, minus the squared deviation around the class means, divided by the total. Zero is useless, one is perfect.



Corn

So you can say these four cuts explain ninety-four percent of the variance.



Herman

Right, and that's the number to put next to your breaks. It also gives you a way to compare class counts. Run it with three classes, then four, then five, and watch how the number climbs. If going from three to four buys you two points and four to five buys you half a point, you've found where the returns flatten.



Corn

Like an elbow plot.



Herman

Same idea. Now, why one dimension is special. Finding the globally best k-means is NP-hard in multiple dimensions. So the usual Lloyd's algorithm gives you a local optimum that depends on the random seed. Run it twice, get different tiers.



Corn

Which is a bad look on a chart you're defending.



Herman

Imagine someone asks why a model sits in tier three, and the honest answer is that it did on Tuesday's run. In one dimension there are polynomial-time algorithms. The kmeans1d library uses dynamic programming, order k times n plus n log n. You get the true optimum. No seeds, no reruns.



Corn

So for a few dozen prices, the answer is deterministic.



Herman

Completely. The tooling is small and boring, which is the point. jenkspy is MIT licensed, Cython, wheels on PyPI and conda-forge. You can call a function with the class count, or use the scikit-learn style object with four classes. One trap. The README example runs it on zero through eleven and gets breaks at zero, two, five, eight, eleven.



Corn

That's five numbers for four classes.



Herman

Because the minimum and maximum are included. The real cuts are the inner breaks, two, five and eight. People plot the wrong lines on that one all the time. Beyond that there's kmeans1d, a ckmeans package wrapping the R implementation, and fast1dkmeans if memory matters, which for Daniel it doesn't.



Corn

How would that trap show up on his chart, in practice?



Herman

He'd draw four vertical lines, and two of them would sit at the very edges of the data. The chart would look like it had five tiers, or the outer lines would just hug the axis and he'd wonder why they're there. It's a quick check. Count the breaks and make sure you have



Herman (continued)

one fewer than the class edges.



Corn

I'd have guessed quantiles were the safe default. Equal counts, no drama.



Herman

That's the common belief, and it's wrong for skewed data. Each class gets the same number of items. Esri's documentation is blunt that the map can be misleading, because similar features land in adjacent classes and widely different values land in the same one.



Corn

So a dense cheap cluster gets sliced in half.



Herman

Picture twenty models priced between fifty cents and a dollar fifty. Quantiles will draw a boundary straight through the middle of them, so a seventy-cent model and a ninety-cent model end up in different tiers. And a two-dollar model and a sixty-dollar model can share a tier. It's the most robust to outliers and the least faithful to groupings. Robust to the tail precisely because it doesn't look at the values.



Corn

It ignores the prices. Bold for a price analysis.



Herman

Then the density approach. A kernel density estimate draws a smooth curve and you cut at the valleys between humps. Conceptually that's closest to "where does the data naturally split." The catch is bandwidth. With a few dozen points, a slightly wider bandwidth erases a valley and a slightly narrower one invents three.



Corn

So you tune the smoothing until it shows the tiers you expected.



Herman

Which is the danger. You can always find a bandwidth that confirms what you already believed. Gaussian mixtures are the standard machinery for choosing the count, you fit several and compare with BIC. I'm not going to cite a particular paper on that, it's just standard practice. But with small samples, mixtures get unstable too.



Corn

Unstable how, in practice?



Herman

Drop one or two models from the list and the fitted components can rearrange themselves. Each component wants enough points to estimate a mean and a spread, and a tier of three expensive models barely gives it that.



Corn

So it's trying to learn a width from three numbers.



Herman

And three numbers can describe almost any width. The fit can shrink that component to a spike on one model, or stretch it to swallow its neighbours, and either one can score well.



Corn

What's the hierarchy one?



Herman

Head and tail breaks, from Bin Jiang, in The Professional Geographer in 2013. Split at the mean. The values above the mean are the head, the minority. Recurse on the head until it's no longer heavy-tailed. The number of classes falls out on its own.



Corn

That sounds built for model pricing. Huge cheap pile, handful of frontier models.



Herman

Exactly that shape. Think of city sizes or word frequencies, where the average is dragged up by a few giants so only a small share of items sit above it. Split, then do it again on the giants. Though it doesn't aim for four. It gives you whatever count the tail structure produces, so if Daniel needs exactly four for a quadrant chart, he may have to merge or stop the recursion early.



Corn

And which wins?



Herman

None, and the literature admits it. A 1986 comparison found none of the traditional methods consistently produced homogeneous classes, each did well only on certain distributions. A 2025 paper makes the same point, that the same data can be presented differently depending on how it's binned. Your own checking of the methods against each other, which I did across a few benchmarking episodes, comes down to the same thing: the metric shapes the conclusion.



Corn

Fine, the methods are settled. But I suspect what you feed them matters more than which one you pick.



Herman

It does. Log scale isn't a method, it's preprocessing, and it changes the answer more than the algorithm. Jenks minimises squared error in whatever units you give it. In raw dollars, the sixty-dollar model dominates the objective and the algorithm spends its effort on the top end.



Corn

So the cheap cluster, where most of the models are, gets lumped together.



Herman

A gap between sixty and a hundred and twenty dollars costs enormous squared error, while the difference between fifty cents and two dollars costs almost nothing. On log ten of price, the squared error tracks ratios. And an order of magnitude more expensive is what a reader actually means by tier. Esri even ships geometrical interval for this, class breaks on a geometric series. Their example is rainfall where most stations record zero.



Corn

Fifteen of a hundred stations record anything, if I remember.



Herman

Yes, I flipped it earlier, eighty-five recording zero is the same thing. A long tail, built for it.



Corn

And a zero in prices would break a log.



Herman

Good catch. Free models, price zero, you either drop them from the fit and make them their own tier, or add a small offset. Decide on purpose, because the size of the offset changes how far the free models sit from the cheap ones on the log axis.



Corn

Make that concrete. What does the offset do?



Herman

Add one cent and a free model sits at log ten of point zero one, which is minus two. Add a tenth of a cent and it sits at minus three. A fifty-cent model is around minus point three either way. So the free models are two units away from the cheap cluster in one case and nearly three in the other, and that gap can decide whether they join the cheap tier or stand alone.



Corn

What about outliers? Daniel specifically doesn't want a few expensive models setting the boundaries.



Herman

Three honest options. Trim, transform, or use a method that ignores the tail. Quantiles ignore the tail by construction. Head and tail breaks quarantines it. The practical question is whether the top few are a real tier or noise.



Corn

My instinct is a handful of frontier models at five hundred dollars is a tier. They're priced differently on purpose.



Herman

Then keep them. I'd say look at the log plot first, and if they sit in a visible gap, they're a tier. If they trail off smoothly, they're a tail.



Corn

So a tail you could cut anywhere, versus a cluster you can't.



Herman

Yes. If you can slide a boundary up or down the tail and nothing visibly changes, the cut is arbitrary and you should say so.



Corn

And then there's what happens when he makes a second chart.



Herman

Esri warns that natural breaks are data-specific, not useful for comparing maps built from different data. If he later charts input prices next to output prices, or batch against standard, the breaks won't line up. A "tier two" on one chart would mean a different price range on the other. For charts that share a scale, quantiles or fixed geometric breaks travel better.



Corn

So a reader flipping between the two charts would think a model moved tier when only the ruler changed.



Herman

And they'd draw conclusions about pricing strategy from what is really an artefact of the binning.



Herman

Class count, too. Every tool treats it as your choice. Jenks designed for fewer than seven because of monochrome shading, so four is well inside the intended use.



Corn

Does that mean the data can't tell him four versus three?



Herman

Not on its own. The fit number can show diminishing returns, but it's a judgment. If the jump from three to four is real but small, the quadrant layout he wants may be reason enough to take four. Just say that's why.



Herman

So the workflow. Plot the sorted values and a density curve on a log axis. Run Jenks or k-means on log prices, and quantiles as a sanity check. Report the fit number. Prefer cuts where both land near each other.



Corn

So the defensible part is the agreement, not the algorithm.



Herman

Yes. And only then hand the four ranges to a language model for names. Name first and you tune the cuts to fit the labels. Also a model will happily name four tiers for data that supports three.



Corn

Well, that is a good place to pause the technical side, I think.

D

Desmond

Before you go on, this is Desmond's Money Tip. Only laugh at jokes that are worth it. I work out the savings in advance, then decide whether the joke has earned a laugh. Roughly forty laughs a year less, at a tenth of a penny of effort each, is four pence saved over twelve months.

D

Desmond

I keep a small tin by the kettle where the laugh money would have gone. It is still empty. That is the system working.

D

Desmond

Money is just energy you have not spent yet, and you will spend it.



Corn

I might try that. Not the tin, the discipline.



Herman

I'm appalled. Four pence.



Corn

You were saying, about the second chart.



Herman

Which one's the problem is when the breaks drift as the prices update. Then the tiers move every month.



Hilbert

Which is my whole complaint. A man at the hardware merchant on Jaffa Road, Mr Abramson, sold me a quarter-inch drill bit in nineteen shillings' worth of grief, I mean forty-two shekels, because he re-sorted his bins every spring. Small, medium, large. I lost two hundred and ten shekels in a year on screws that moved shelves.



Hilbert

Herman, I rule my own chart in pencil. Four bands. The boundaries have not moved since I drew them, and the prices underneath have wandered a long way. If a new cut lands within a pencil-width of the old line, it stays. I don't need a library for that. I re-binned once. I'm not saying what.



Herman

A pencil-width is a tolerance. It's a hysteresis band, honestly, which is a real technique for stopping boundaries flickering. Thermostats do it. The heating doesn't switch on at nineteen degrees and off at nineteen and a bit, it waits for a gap, so it doesn't chatter.



Corn

Though a pencil gets shorter. So the tolerance shrinks as you sharpen it.



Hilbert

I use a mechanical one. Point five millimetre. Fine, that's a correction, Corn, I take it.



Herman

That actually connects to the second-chart problem. Stable cuts you only move when the data moves them past a tolerance.



Corn

How would that work for Daniel monthly?



Herman

Rerun the cuts each month on log prices. For each new break, compare it to the published one. If it's inside the tolerance, say a few percent in price terms, keep the old line. If it's outside, move it and note the change.



Corn

And if a price drifts but stays inside the band?



Herman

Then the model keeps its tier and nobody sees a change. The only moves readers notice are the ones that actually mean something, like a provider cutting its price by half and crossing a line.



Corn

Then the real question for Daniel is what the chart is for. Comparing models to each other wants portable cuts. Showing the shape of one market wants data-specific ones.



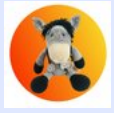
Herman

And since no method wins on its own, the agreement check is what he can actually defend. Keep a tolerance on the old lines, as we just heard, so the tiers don't wobble each time a price changes.



Corn

Thanks to Hilbert Flumingtop for producing, as always.



Herman

If you enjoyed this one, try episode nine, Benchmarking Custom ASR Tools - Beyond The WER; episode seven, Building Custom ASR Tools; and episode twelve, The AI Breakthrough. This has been My Weird Prompts. If you enjoyed it, a review helps other people find the show.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Esri ArcGIS Pro documentation, current [primary]**
https://doc.esri.com/en/arcgis-pro/latest/help/mapping/layer-properties/data-classification-methods.html

2. **mthh/jenkspy, MIT, release 0.4.1 (2024-06-03), last push 2026-02-13 [primary]**
https://github.com/mthh/jenkspy

3. **dstein64/kmeans1d, MIT [primary]**
https://github.com/dstein64/kmeans1d

4. **Wikipedia, last edited 2026-10-20 [primary for method definition]**
https://en.wikipedia.org/wiki/Jenks_natural_breaks_optimization

5. **Wikipedia, last edited 2026-10-19**
https://en.wikipedia.org/wiki/K-means_clustering

6. **Daniel Steinberg, 2019-07-08**
https://www.dannyadam.com/blog/2019/07/kmeans1d-globally-optimal-efficient-1d-k-means/

7. **PyPI v0.5.0**
https://pypi.org/project/kmeans1d/

8. **PyPI v1.2.0 (Ckmeans.1d.dp wrapper)**
https://pypi.org/project/ckmeans/

9. **PyPI v0.1.2**
https://pypi.org/project/fast1dkmeans/

10. **Bin Jiang, Head/tail Breaks, submitted 2012-09-13; *The Professional Geographer* 65(3), 2013, 482**
494
https://arxiv.org/abs/1209.2801v1

11. **Comparing traditional methods for selecting class intervals on choropleth maps, *The Professional Geographer*, 1986**
https://doi.org/10.1111/j.0033-0124.1986.00062.x

12. **Exploropleth: exploratory analysis of data binning methods in choropleth maps, *Cartography and Geographic Information Science*, 2025**
https://doi.org/10.1080/15230406.2025.2543483

13. **Rust crate, Jenks and quantile breaks**
https://github.com/Nik4002/classify

14. **Rust, MIT, 2026**
https://github.com/uratne/natural-breaks